DOCUMENT RESUME

ED 307 298 TM 013 315

AUTHOR Crehan, Kevin; Haladyna, Thomas M.
TITLE The Validity of Two Item-Writing Rules.
PUB DATE 89
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Higher Education; *Multiple Choice Tests; Psychology; *Test Construction; Test Format; Test Validity
IDENTIFIERS *Item Writing Rules; Parallel Test Forms; Stem Analysis

ABSTRACT

The present study involved the testing of two common multiple-choice item writing rules. A recent review of research revealed that much of the advice given for writing multiple-choice test items is based on experience and wisdom rather than on empirical research. The rules assessed in this study include: (1) the phrasing of the stem in the form of a question versus a partial sentence; and (2) the use of the inclusive "none of the above" option instead of a specific content option. Limited empirical research suggests that using the partial sentence format and the inclusive "none of the above" option may lead to undesirable item and test characteristics, while textbook authors essentially are divided on their opinions about the validity of each rule. The items used in this study were from the instructor's manual for D. Myer's (1986) text entitled "Psychology." Items were randomly assigned to be rewritten to reflect the experimental conditions under investigation. Two instructors of an introductory psychology course selected 32 multiple-choice items for the study. The rewritten tests were administered to 228 students enrolled in two sections of an introductory psychology class. About half of the students in each section received Form A and the other half received Form B, resulting in 115 Form A and 113 Form B responses. The same manipulated items were combined with 18 different non-manipulated items in a third section of the class to comprise Forms C and D, whose administration resulted in 59 Form C and 59 Form D responses. Results offer no evidence to support the use of either type of stem and limited evidence to caution against use of the "none of the above" option. Two data tables and examples of the four item formats used are provided. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307298

The Validity of Two Item-Writing Rules

Kevin Crehan

University of Nevada, Las Vegas

and

Thomas M. Haladyna

Arizona State University West Campus

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KEVIN D. CREHAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM013315

BEST COPY AVAILABLE

The Validity of Two Item-Writing Rules

ABSTRACT

A recent review of research revealed that much of the advice given for writing multiple-choice test items is based on experience and wisdom rather than empirical research. The present study involved the testing of two common item writing rules: (1) the phrasing of the stem in the form of a question versus a partial sentence and (2) the use of the inclusive "none of the above" option instead of a specific content option. Limited empirical research suggests that using the partial sentence format and the inclusive 'none of these' option may lead to undesirable item and test characteristics, while textbook authors essentially are divided on their opinions about the validity of each rule. Results of this experimental study offer no evidence to support the use of either type of stem and limited evidence to caution against use the option "none of the above."

The Validity of Two Item-Writing Rules

A number of writers in the field of educational measurement have commented that multiple-choice (MC) item writing, despite its widespread popularity and use, has received little scholarly attention in the past (Cronbach, 1970; Ebel, 1951; Millman & Green, in press; Nitko, 1984; Roid and Haladyna, 1982; Wesman, 1971; Wood, 1977). In a review of empirical research on item writing, Haladyna and Downing (1989a) reported finding 96 empirical studies of which 53 dealt with only two item-writing practices, the optimal number of options and the desirability of key balancing. Most item-writing rules have been studied fewer than 10 times. Thus the empirical foundation for the validity of many item-writing rules is weak, and the basis for many rules is often authoritative wisdom passed on through textbooks and other professional publications and presentations.

The study reported here addresses two item-writing rules which are popularly prescribed in treatments on MC item writing in textbooks and other sources in the educational measurement literature (Haladyna & Downing, 1989b). The first rule is: "Don't use 'none of the above' as an option"; the second rule is "Use either the question format or the completion format when phrasing the stem."

None of the Above

In a review of 46 references dealing with the topic MC

item writing, Haladyna & Downing (1989b) found that 34 (73%) of these references stated support or lack of support for the "Don't use 'none of the above' as an option" rule. This was the tenth most often mentioned rule, and this survey was taken as evidence of the importance of the rule for item writers. However, authors were divided on their support for this rule, with 19 for and 15 against. Obviously some controversy exists in the validity of the rule.

Empirical research on this item writing rule has been limited to only ten studies (Boynton, 1950; Dudycha & Carpenter, 1973; Forsyth & Spratt, 1980; Hughes & Trimble, 1965; Mueller, 1975; Oosterhof & Coats, 1984; Rimland, 1960; Schmeiser & Whitney, 1975; Wesman & Bennett, 1946; Williamson & Hopkins, 1967). All of these studies involved the item characteristic of difficulty, but only five studied item discrimination and reliability, and only two validity. In all instances, the use of "none of the above" option made items more difficult, the mean effect across nine studies where results were aggregable was 4.8%. With discrimination, avoiding the inclusive "none of the above" option made items slightly more discriminating, .03, while reliability was improved by a factor of .04.

Question Format Versus Completion Format

One of the most fundamental requirements in MC item writing is that one states the item in a question format or a completion format. On the surface there appears to be no reason

to challenge either format. According to Haladyna & Downing (1989a), the rule is one of the most common given in treatments on MC item writing, 41 of 46 references mentioned it, and all 41 support the use of either format. Paradoxically, the small body of empirical research leads to the opposite conclusion.

Studies of this item writing rule include: Board and Whitney (1972), Dudycha & Carpenter (1973), Dunn & Goldstein (1959), Schmeiser & Whitney (1975a; 1975b), and Schrock & Mueller (1982). These six studies observed effects on item difficulty in each instance, discrimination in three cases, reliability four times, and validity twice. In general, the question format appears to have an advantage over the sentence completion format with respect to making items slightly easier, having little or no effect on item discrimination, and making test scores based on such items more reliable and valid. For reliability, the improvement was a median .065, which is a reduction of 6.5% error variance in test scores. Validity was improved by .06 in two studies (Board & Whitney, 1972; Schmeiser & Whitney, 1975b). Based on these few studies, it appears the evidence favors the use of the question format over the completion format in phrasing the MC stem.

The present study further investigates these two item-writing rules.

METHOD

The items used in this study were from the instructor's manual for Myer's (1986) text entitled Psychology. Two instructors of an introductory psychology course selected 32 MC items for the study. Each item was keyed to the objectives of the course and met the standard requirements for MC item writing. Each item also had adequate performance characteristics as judged from previous uses. Items were randomly assigned to be rewritten to reflect the experimental manipulations as outlined below:

No. of items	Version 1	Version 2
8	completion option 'e' (CE)	completion none of these (CN)
8	question option 'e' (QE)	completion option 'e' (CE)
8	question none of these (QN)	question option 'e' (QE)
8	completion none of these (CN)	question none of these (QN)

Figure 1 provides an example of one item written in all four variations.

Insert Figure Inabout here

The manipulations were balanced both within and between the two versions. Version 1 items were combined with eighteen non-manipulated items to comprise Form A of the final exam for two sections of an introductory psychology class while Version 2 items were combined with the same eighteen items to comprise Form B. Test forms were key balanced with the option 'none of these' being keyed three times in sixteen appearances or approximately one-fifth of the time.

The tests were administered to two sections of the class with approximately one-half the students in each section receiving Form A and the other half receiving Form B resulting in 115 Form A and 113 Form B responses. In addition, the same manipulated items were combined with eighteen different non-manipulated items in a third section of the class to comprise Forms C and D. Forms were key balanced as above and test administration in this class resulted in 59 Form C and 59 Form D responses.

This design was chosen to allow comparison of item format manipulations controlling for examinee ability. That is, when Version 1 CE items are combined with Version 2 QE items, we have sixteen items not employing the option 'none of these'. When Version 1 QN items are combined with Version 2 CN items we have these same sixteen items employing the option 'none of these'. Item characteristics can be compared between these sixteen item

sets since all subjects in the study responded to one or the other of the eight item subscales under each condition. Since, at best, small effect sizes were anticipated hypothesis testing was conducted with alpha set at the .10 level for each statistical test.

RESULTS

Table 1 presents the means and standard deviations of item difficulties, mean point-biserials and the Kuder-Richardson 20 reliability estimates of each subscale for the four forms of the test.

Insert table 1 about here

In order to test for differences in difficulty and discrimination for the question versus completion format item statistics for the Form A-QE items were combined with item statistics for the Form B-QN items and were compared to the Form A-CN items combined with the Form B-CE items. Similarly item statistics for the same item types on Forms C and D were combined. In order to test for differences in difficulty and discrimination for the inclusive versus specific option hypothesis item statistics for Form A-CE items were combined with Form B-QE and were compared to the Form A-QN items combined with

Form B-CN items. Similarly, item statistics for the same item types were combined on Forms C and D. Summary statistics for the combined items are presented in Table 2.

Insert Table 2 about here

DIFFICULTY

The observed difference in difficulty was .02 higher for the question format. A correlated one-tailed t-test showed non significance at the .10 level ($t = .56$, $df = 15$, $r = .70$, $p = .29$). The t-test for the same comparison on Forms C and D showed similar results with a mean difference of .003 and a non-significant t statistic ($t=.10$, $df = 15$, $r = .76$, $p = .46$). Differences between using and not using the option 'none of these' was tested by combining Form A CE with Form B QE item difficulties and comparing these with Form A Q and Form B C item difficulties. The difference in mean difficulty was .027 with use of 'none of these' being lower. The dependent t-test was significant at the .1 level ($t = 1.44$, $df = 15$, $r = .916$, $p = .085$). The same test for Forms C and D had similar results with a mean difference of .043 ($t = 1.59$, $df = 15$, $r = .67$, $p = .065$).

DISCRIMINATION

Differences in mean point-biserials between the question and

completion formats were non-significant for both replications. Differences in mean point-biserials between using and not using the inclusive 'none of these' option were .034 and .033 for Form A vs Form B and Form C vs Form D respectively and favored not using the inclusive option in both instances. The observed differences, however, failed to reach significance at the .10 level. The correlated t-tests for Form A versus Form B and Form C versus Form D had p values of .18 and .20 respectively.

DISCUSSION

While this study fails to offer support to a recommendation regarding use of either the question or completion format over the other, observed results regarding use of the "none of these" option are consistent with previous findings in direction and magnitude. Differences in difficulty were statistically significant and in 3 to 4% range favoring the specific option over the inclusive option format. Item discriminations were also observed to be slightly over .033 higher for the specific option format. This result, while not statistically significant, is at the same level as observed in previous research. Lack of statistical significance may be attributable to the low power to detect a difference of this magnitude with sixteen subjects (items) and the low correlations between the item discriminations between forms (.183, .488). It is noted that differences in item

discrimination observed in this study are estimated to result in differences in reliability of about .04 favoring use of the specific option over use of "none of these". Future research on this should use the knowledge of this effect size to determine the sample size necessary to detect a .03 or greater effect with reasonable power.

REFERENCES

1. Board, C. & Whitney, D. R. (1972). The effect of selected poor item- writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 225-233.

2. Boynton, M. (1950). Inclusion of "none of these" makes spelling items more difficult. Educational & Psychological Measurement, 10, 431-432.

3. Cronbach, L. J. (1970). Review of On the theory of achievement test items by J.R. Bormuth. Psychometrika, 35, 509-511.

4. Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as a function of a selected multiple-choice item construction principles. Educational and Psychological Measurement, 19, 171-179.

5. Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. Journal of Applied Psychology, 58, 116-121.

6. Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.) Educational Measurement, (1st ed.) (pp. 185-249). Washington, DC: American Council on Education.

7. Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for test tests. Psychometrika, 34, 363-373.

8. Forsyth, R. A. & Spratt, K. F. (1980). Measuring problem solving ability in mathematics with multiple choice items. The effect of item format on selected item and text characteristics. Journal of Educational Measurement, 17, 31-43.

9. Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education, 3rd ed., New York, N.Y., McGrav-Hill.

10. Haladyna, T. M. & Downing, S. M. (1989a). The validity of a taxonomy of multiple-choice item writing rules. Applied Measurement in Education, 1.

11. Haladyna, T. M. & Downing, S. M. (1989b). A taxonomy of multiple-choice item writing rules. Applied Measurement in Education, 1.

12. Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple-choice items. Educational and

Psychological Measurement, 25, 117-126.

13. Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. Educational and Psychological Measurement, 35, 135-141.

14. Millman, J., & Greene, J. (In press). The specification and development of tests of achievement and abilities. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). Washington, DC: American Council on Education.

15. Myers, D. (1986). Psychology. New York, N.Y.: Worth Publishers.

16. Nitko, A. J. (1984). Book review of Roid and Haladyna's A technology for test-item writing. Journal of Educational Measurement, 21, 201-204.

17. Oosterhof, A. C. & Coats, P. K. (1984). Comparison of difficulties and reliability of quantitative word problems in completion and multiple-choice item formats. Applied Psychological Measurement, 8, 287-294.

18. Rimland, B. (1960). The effects of varying time limits and of using "right answer not given" in experimental forms of the U. S. Navy Arithmetic Test. Educational and Psychological Measurement, 20, 533-539.

19. Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York, NY: Academic Press.

20. Schmeiser, C. B., & Whitney, D. R. (1975a). Effect of two selected item-writing practices on test difficulty, discrimination, and reliability. Journal of Experimental Education, 43, 30-34.

21. Schmeiser, C. B., & Whitney, D. R. (1975b). The effect of incomplete stems and "none of the above" foils on test and item characteristics. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

22. Schrock, T. J., and Mueller, D. J. (1982). Effects of violating three multiple-choice item construction principles. The Journal of Educational Research, 75, 314-318.

23. Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), Educational Measurement (pp.99-111). Washington, DC: American Council on Education.

24. Wesman, A. G., & Bennett, G. K. (1946). The use of 'none

of these' as an option in test construction. Journal of Educational Psychology, 37, 541-549.

25. Williamson, M. L., & Hopkins, K. D. (1967). The use of "none-of-these" versus homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. Journal of Educational Measurement, 4, 53-58.

26. Wood, R. (1977). Multiple choice: A state of the art report. Evaluation in Education: International Progress, 1, 191-280.

TABLE 1

Mean (P) and standard deviation (S) of difficulty indices, mean point-biserials (D) and KR20 reliability (r) for each 8 item subscale across test forms.

Statistic	Item Type	Form A	Item Type	Form B	Item Type	Form C	Item Type	Form D
P	CE	.705	CN	.636	CE	.733	CN	.634
S		.180		.209		.136		.132
D		.379		.401		.450		.386
r		.535		.591		.619		.627
P	QE	.792	CE	.798	QE	.798	CE	.778
S		.140		.130		.149		.143
D		.352		.415		.355		.405
r		.469		.568		.528		.566
P	QN	.806	QE	.790	QN	.731	QE	.718
S		.092		.087		.134		.131
D		.319		.409		.401		.402
r		.388		.489		.611		.572
P	CN	.622	QN	.666	CN	.667	QN	.653
S		.217		.158		.202		.192
D		.328		.419		.395		.415
r		.385		.580		.549		.539

TABLE 2

Means and standard deviations for item difficulties and discriminations, and with estimated reliability on the combined sixteen item scales for each item type

Forms	Item Type	Mean Diff.	Standard Deviation	Mean Disc.	Standard Deviation	Reliability*
A&B	Q	.729	.159	.386	.114	.74
A&B	C	.710	.195	.371	.106	.72
C&D	Q	.725	.181	.384	.124	.74
C&D	C	.722	.178	.400	.142	.75
A&B	E	.748	.144	.394	.107	.75
A&B	N	.720	.179	.360	.117	.70
C&D	E	.726	.129	.426	.168	.78
C&D	N	.682	.138	.393	.125	.75

*Reliability estimate based on average point-biserials for sixteen items after Guilford (1965).

Figure 1

The following is an example of the four item formats appearing on the criterion instruments.

(CN) In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to:

a. stomach ulcers.
b. cancer.
* c. heart attacks.
d. accidents.
e. none of these

(QN) In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to which of the following?

a. stomach ulcers
b. cancer
c. strokes
d. accidents
* e. none of these

(CE) In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to:

a. stomach ulcers.
b. cancer.
* c. heart attacks.
d. accidents.
e. strokes.

(QE) In their classic nine-year study, Friedman and Roseman found that competitive, hard-driving, impatient, and easily angered individuals are especially susceptible to which of the following?

a. stomach ulcers
b. cancer
* c. heart attacks
d. accidents
e. strokes